What is industry Looking for Analytics Hires? Text Mining of Placement Ads

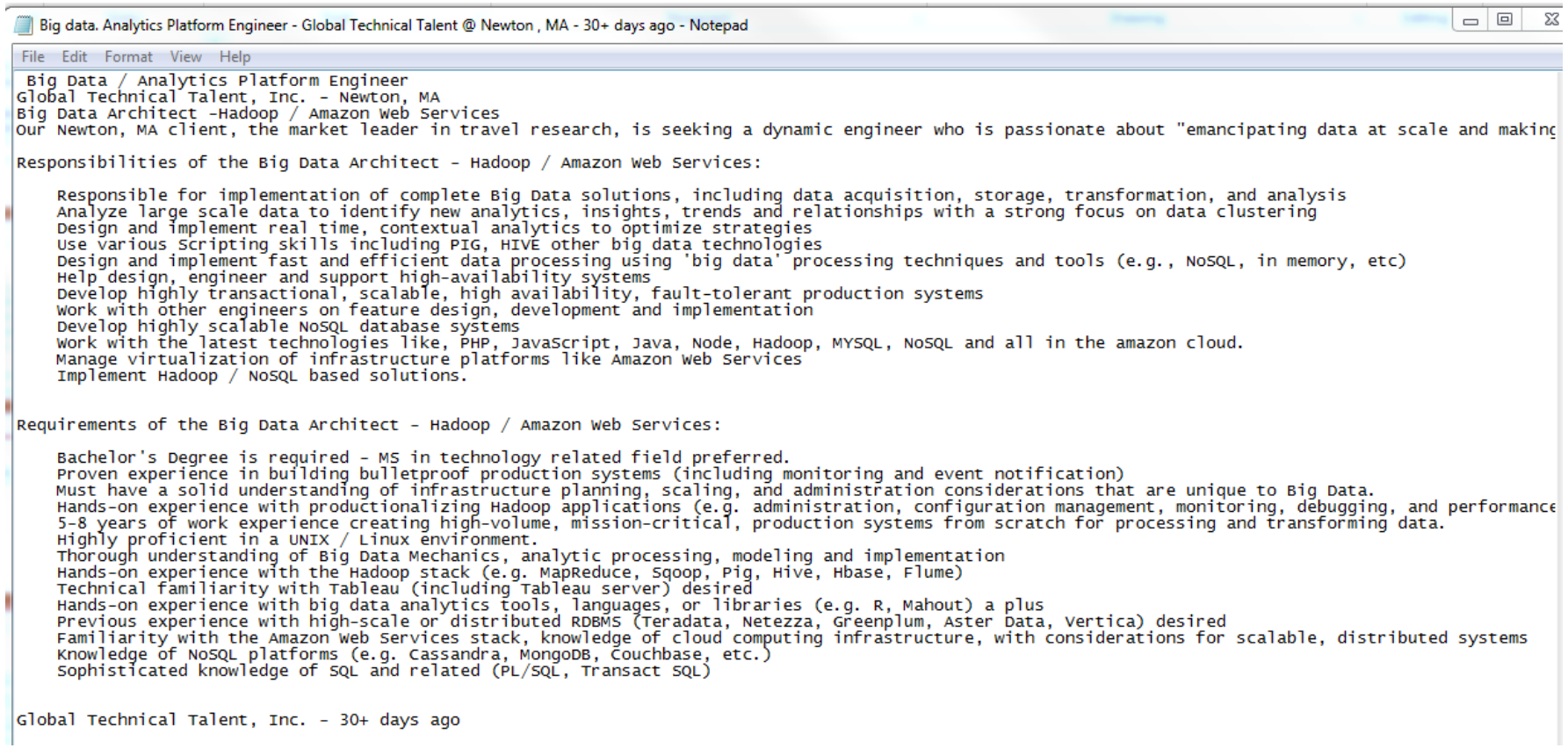
Dr. Goutam Chakraborty, Ralph A. and Peggy A. Brenneman Professor (Marketing)
Hamed M. Zolbanin, Ph.D. Candidate (MSIS)

INFORMS 2014

Data Collection

- Data collected between March and July 2014
- 724 job postings on LinkedIn (53%), iCrunchData (21%), Indeed(15%) and Monster(11%) focused on US jobs
- Keywords used: “Big data”, “Analytics”, “Business Analyst” and “Data scientist”
- Data captured as a text file for each job: Job title, date of posting, company name, company location, job requirements

Example Text File



Big data. Analytics Platform Engineer - Global Technical Talent @ Newton , MA - 30+ days ago - Notepad

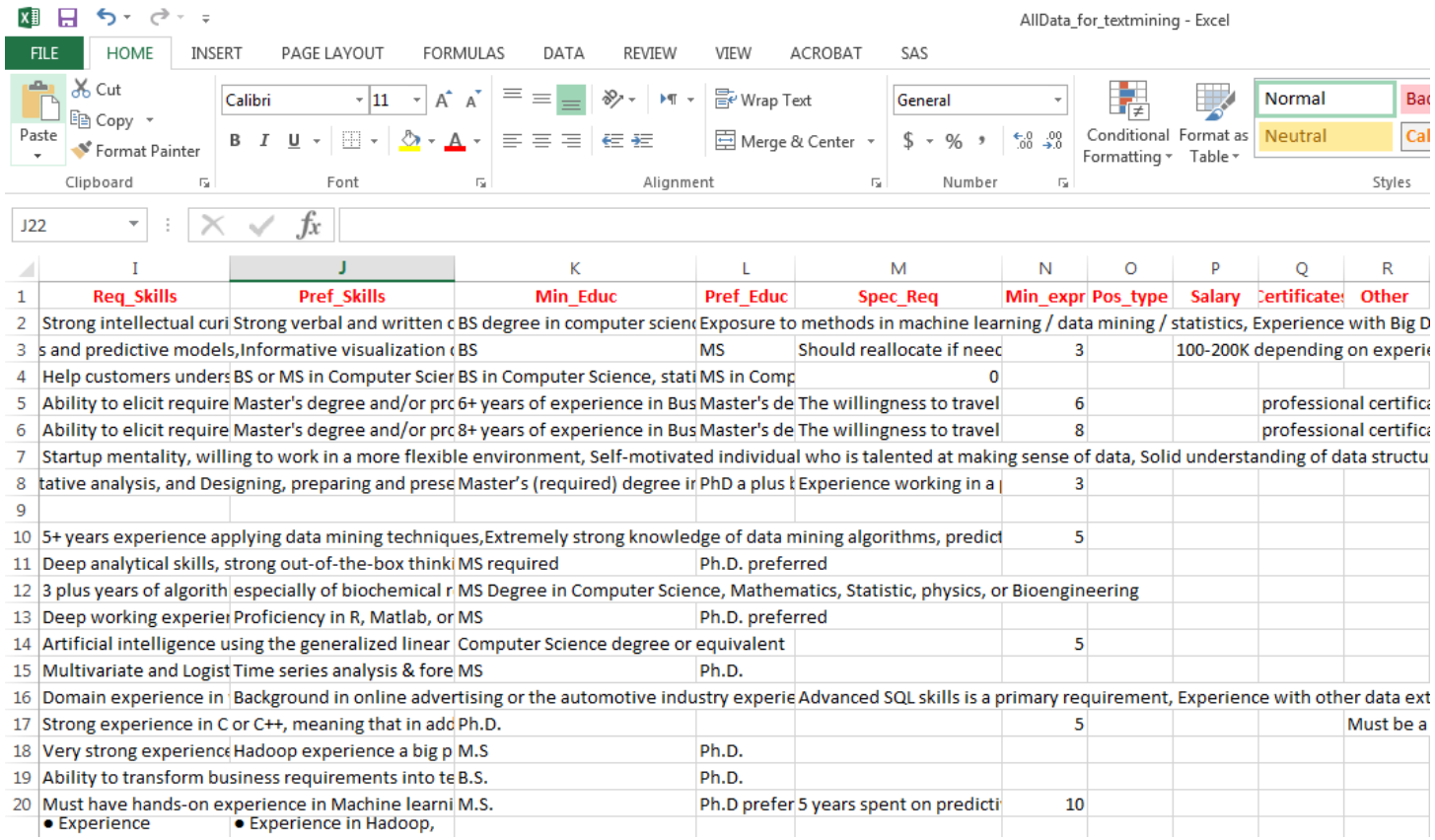
File Edit Format View Help

```
 Big Data / Analytics Platform Engineer
Global Technical Talent, Inc. - Newton, MA
Big Data Architect -Hadoop / Amazon Web Services
Our Newton, MA client, the market leader in travel research, is seeking a dynamic engineer who is passionate about "emancipating data at scale and making

Responsibilities of the Big Data Architect - Hadoop / Amazon Web Services:

    Responsible for implementation of complete Big Data solutions, including data acquisition, storage, transformation, and analysis
    Analyze large scale data to identify new analytics, insights, trends and relationships with a strong focus on data clustering
    Design and implement real time, contextual analytics to optimize strategies
    Use various Scripting skills including PIG, HIVE other big data technologies
    Design and implement fast and efficient data processing using 'big data' processing techniques and tools (e.g., NoSQL, in memory, etc)
    Help design, engineer and support high-availability systems
    Develop highly transactional, scalable, high availability, fault-tolerant production systems
    Work with other engineers on feature design, development and implementation
    Develop highly scalable NoSQL database systems
    Work with the latest technologies like, PHP, JavaScript, Java, Node, Hadoop, MYSQL, NoSQL and all in the amazon cloud.
    Manage virtualization of infrastructure platforms like Amazon Web Services
    Implement Hadoop / NoSQL based solutions.


Requirements of the Big Data Architect - Hadoop / Amazon Web Services:

    Bachelor's Degree is required - MS in technology related field preferred.
    Proven experience in building bulletproof production systems (including monitoring and event notification)
    Must have a solid understanding of infrastructure planning, scaling, and administration considerations that are unique to Big Data.
    Hands-on experience with productionalizing Hadoop applications (e.g. administration, configuration management, monitoring, debugging, and performance
    5-8 years of work experience creating high-volume, mission-critical, production systems from scratch for processing and transforming data.
    Highly proficient in a UNIX / Linux environment.
    Thorough understanding of Big Data Mechanics, analytic processing, modeling and implementation
    Hands-on experience with the Hadoop stack (e.g. MapReduce, Sqoop, Pig, Hive, Hbase, Flume)
    Technical familiarity with Tableau (including Tableau server) desired
    Hands-on experience with big data analytics tools, languages, or libraries (e.g. R, Mahout) a plus
    Previous experience with high-scale or distributed RDBMS (Teradata, Netezza, Greenplum, Aster Data, Vertica) desired
    Familiarity with the Amazon Web Services stack, knowledge of cloud computing infrastructure, with considerations for scalable, distributed systems
    Knowledge of NoSQL platforms (e.g. Cassandra, MongoDB, Couchbase, etc.)
    Sophisticated knowledge of SQL and related (PL/SQL, Transact SQL)


Global Technical Talent, Inc. - 30+ days ago
```

Data Cleaning

- Lack of a standard for job postings within and across the websites
- The collected data were manually entered into an MS Excel document under different columns as specified by the researchers
- Several aspects of the data were captured by designation of different columns for the collected data
- More effective and informative analysis by separating data into multiple columns
  - "job title", "job summary", "**required skills**", "minimum education", "minimum experience", etc.

Example of Excel File



	I	J	K	L	M	N	O	P	Q	R
1	Req_Skills	Pref_Skills	Min_Educ	Pref_Educ	Spec_Req	Min_expr	Pos_type	Salary	Certificate	Other
2	Strong intellectual curi	Strong verbal and written c	BS degree in computer scienc	Exposure to methods in machine learning / data mining / statistics, Experience with Big D						
3	s and predictive models,Informative visualization (		BS	MS	Should reallocate if neec	3		100-200K depending on experie		
4	Help customers unders	BS or MS in Computer Scier	BS in Computer Science, stati	MS in Comp	0					
5	Ability to elicit require	Master's degree and/or pro	6+ years of experience in Bus	Master's de	The willingness to travel	6			professional certifica	
6	Ability to elicit require	Master's degree and/or pro	8+ years of experience in Bus	Master's de	The willingness to travel	8			professional certifica	
7	Startup mentality, willing to work in a more flexible environment, Self-motivated individual who is talented at making sense of data, Solid understanding of data structu									
8	tative analysis, and Designing, preparing and prese		Master's (required) degree ir	PhD a plus k	Experience working in a	3				
9										
10	5+ years experience applying data mining techniques,Extremely strong knowledge of data mining algorithms, predict					5				
11	Deep analytical skills, strong out-of-the-box thinki		MS required	Ph.D. preferred						
12	3 plus years of algorith	especially of biochemical r	MS Degree in Computer Science, Mathematics, Statistic, physics, or Bioengineering							
13	Deep working experier	Proficiency in R, Matlab, or	MS	Ph.D. preferred						
14	Artificial intelligence using the generalized linear		Computer Science degree or equivalent			5				
15	Multivariate and Logist	Time series analysis & fore	MS	Ph.D.						
16	Domain experience in	Background in online advertising or the automotive industry experie			Advanced SQL skills is a primary requirement, Experience with other data ext					
17	Strong experience in C or C++, meaning that in add		Ph.D.			5				Must be a
18	Very strong experience	Hadoop experience a big p	M.S	Ph.D.						
19	Ability to transform business requirements into te		B.S.	Ph.D.						
20	Must have hands-on experience in Machine learni		M.S.	Ph.D prefer	5 years spent on predicti	10				
	• Experience	• Experience in Hadoop,								

SAS Text Mining Options For the Variable "Required Skills"

- Spell checking set to "YES"
- Standard dictionary used in "Text Filter" node
- Frequency weighting set to "Default"
- Term Weight set to "Default"
- Minimum number of documents: 4 (excludes terms that appear in less than 4 documents)
- A synonym list created for the terms appeared in documents; e.g., Support Vector Machines and SVM were considered synonyms (see next slide)

Example of Synonyms

curious mind	NOUN_GROUP	curiosity	Noun
data accuracy	NOUN_GROUP	data management	NOUN_GROUP
data acquisition	NOUN_GROUP	data management	NOUN_GROUP
data analysis	NOUN_GROUP	data management	NOUN_GROUP
data analytics	NOUN_GROUP	data management	NOUN_GROUP
data architecture	NOUN_GROUP	data management	NOUN_GROUP
data collection	NOUN_GROUP	data management	NOUN_GROUP
data extraction	NOUN_GROUP	data management	NOUN_GROUP
data integration	NOUN_GROUP	data management	NOUN_GROUP
data manipulation	NOUN_GROUP	data management	NOUN_GROUP
data mine	NOUN_GROUP	data management	NOUN_GROUP
data optimization	NOUN_GROUP	data management	NOUN_GROUP
data preparation	NOUN_GROUP	data management	NOUN_GROUP
data process	NOUN_GROUP	data management	NOUN_GROUP
data quality	NOUN_GROUP	data management	NOUN_GROUP
data science	NOUN_GROUP	data management	NOUN_GROUP
data solution	NOUN_GROUP	data management	NOUN_GROUP
data structure	NOUN_GROUP	data management	NOUN_GROUP
data system	NOUN_GROUP	data management	NOUN_GROUP
data validation	NOUN_GROUP	data management	NOUN_GROUP
data visualization	NOUN_GROUP	data management	NOUN_GROUP
data warehouse	NOUN_GROUP	data management	NOUN_GROUP
database	Noun	database technol...	NOUN_GROUP
database design	NOUN_GROUP	database technol...	NOUN_GROUP

Initial Results (Clusters)

- 17+1 different clusters for jobs based on their requirements (one cluster for jobs without required skills – it had other fields such as job title, job summary..)

Cluster ID	Descriptive Terms		Frequency	Percentage
1			49	7%
2	+model +statistical +'cluster analysis' +'decision tree' +'sas institute' +'statistical model' +leadership +decision +technique +research	...	64	9%
3	+discipline +problem +'excellent communication' +test +technique +machine +design +'core business problem' +'decision tree' +linux	...	16	2%
4	+english +fluent +'bachelor's degree' degree travel +year 'business intelligence' +discipline +design +project	...	22	3%
5	+nosql +pig +oracle +hive spss +hadoop +'database technology' +'relational database' sql 'business intelligence'	...	53	7%
6	+programming +'algorithm development' +language +system +'machine learn technique' +code +machine +python +java +nosql	...	45	6%
7	+assessment +conduct +theory +'text analytics' +principle +'predictive analytics' +practice +assess +product +research	...	7	1%
8	excel microsoft powerpoint +skill +excellent +detail verbal multiple written +ability	...	89	12%
9	+python computer +machine learning +'algorithm development' +language +'machine learn technique' +java engineering +programming	...	67	9%
10	+track +record proven complex +solution +problem technical +'functional requirement' +'analytical capability' +solver	...	41	6%
11	'actionable manner' +'high-dimensionality data' +high-volume +manipulate +vary precise comfort +'complex quantitative analysis' complex actionable	...	11	2%
12	+enhance +assess +manipulation +function +articulate +oracle +learn technical +'functional requirement' +methodology	...	13	2%
13	+'core business problem' +challenge +'natural language' +level +insight +core +theory +'machine learn technique' +code +articulate	...	7	1%
14	+'decision tree' +network +regression +'cluster analysis' +'algorithm development' +'natural language' +decision +'machine learn technique' +machine +'text analytics'	...	29	4%
15	+year +'sas institute' +model +experience engineering +manipulate +amount travel +articulate spss	...	18	2%
16	'google analytics' +adobe +marketing +analytics +test +advanced +tool excel +report +prefer	...	59	8%
17	+project +process +business +team +manage +report +solution actionable +practice +test	...	111	15%
18	+linux +script +language verbal +'written communication skill' +python proven +excellent +network working	...	25	3%

Initial Results (Topics)

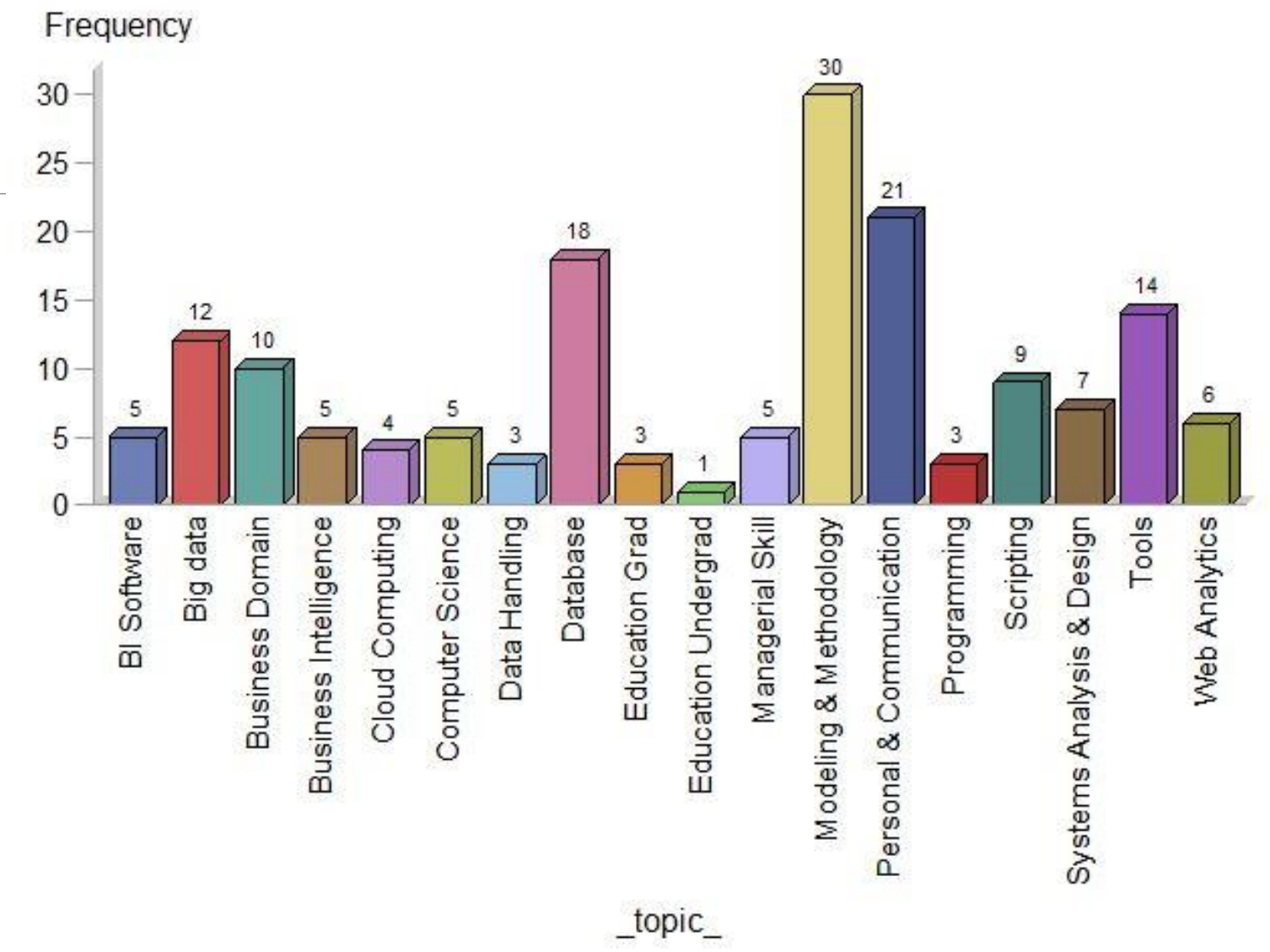
- 25 default topics identified by SAS Text Mining facilities

Topic ID ▲	Topic
1	+teradata,competencies,preferable experience,preferable,quickly
2	+report,+identify,data,+report,technical
3	+teradata,hurdle,ambiguity,subject,+seek
4	redpoint,ssis,db2,rapidminer,tibco
5	data,+sas institute,+python,+year,science
6	actionable,+allow,+vary,+manipulate,precision
7	powerpoint,microsoft,+sas institute,data,+word
8	+marketing,+measurement,+website,+report,+report
9	+conversion,+traffic,+site,+website,+assist
10	attention,+problem,+excellent,technical,+ability
11	+compute,+java,data,+supervision,lucene
12	+problem,+require,+python,+solver,data
13	+sas institute,time series,+response,real time analytics,practical knowledge
14	usage,+sas institute,+line,+python,command
15	+sas institute,data,predictive,+assess,+data management
16	+evaluate,+analyze,+site,customer behavior,+success
17	+hierarchical,+python,+cluster analysis,+java,+regression analysis
18	+require,db2,travel,+industrial,+oracle
19	media industry,journalism,+improvement,data,+core business problem
20	data,+oracle,+data management,+analyze,+report
21	+analyze,+leverage,+sas institute,+analytic model,+pattern recognition
22	+analyze,+website,+page,+collect,+duty
23	+metrics,+improvement,+sas institute,+oracle,intellectual
24	+rollout,+recommender system,+improve,+search,+strong interpersonal skill
25	data,quantify,actual,+high level,+success

Customized Topic Creation

- Based on the terms appeared after filtering the texts, a preliminary list of customized topics developed by the research team
- The topic list went through a series of modifications until consensus was achieved
- 18 customized topics
- Each topic consisted of 1 to 30 terms

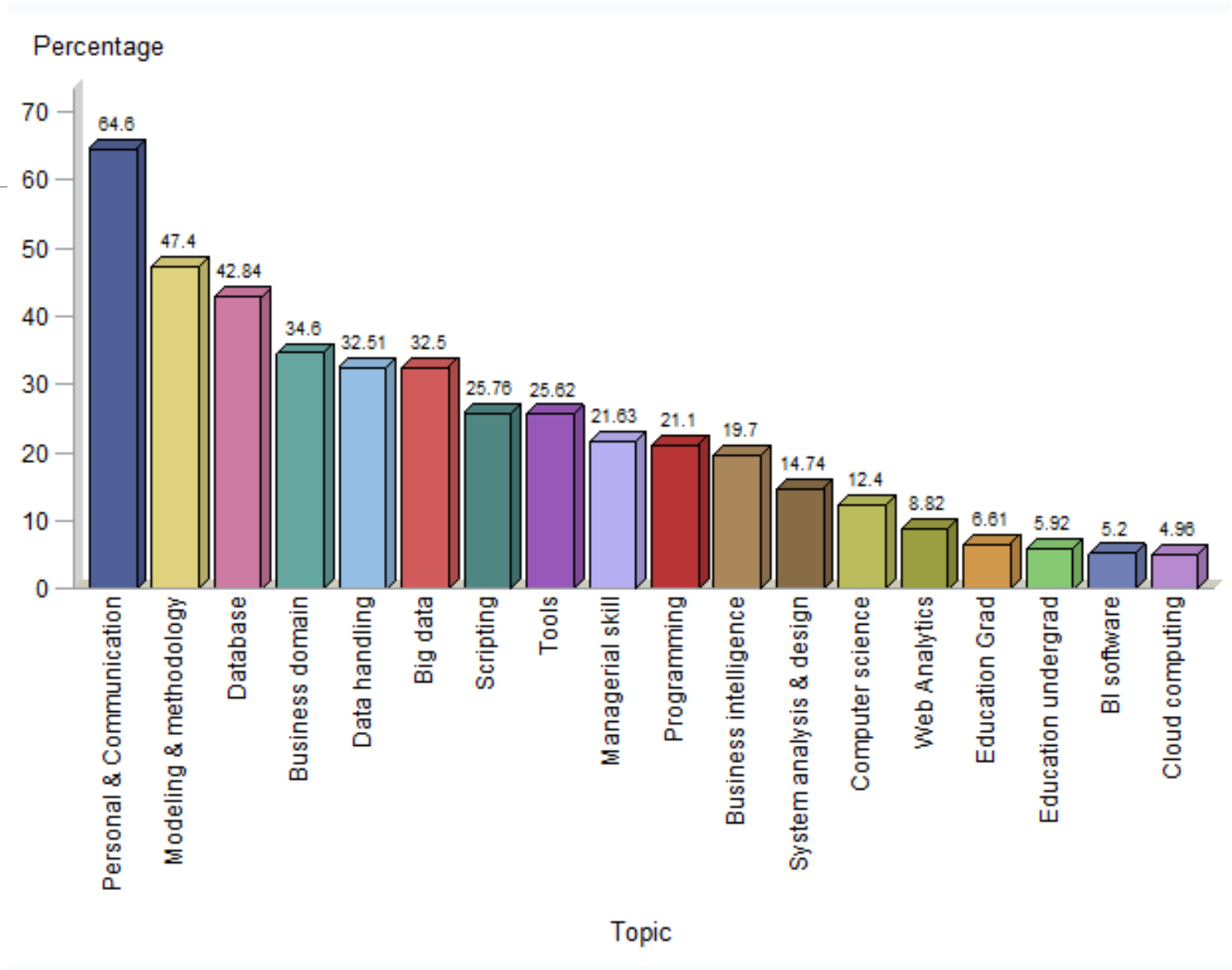
Terms in Each Topic



Customized Terms

topic	term
BI Software	hyperion
BI Software	microstrategy
BI Software	pentaho
BI Software	qlikview
BI Software	ssrs
Big data	big data
Big data	cassandra
Big data	hDFS
Big data	hadoop
Big data	hive
Big data	mahout
Big data	map-reduce
Big data	massive data
Big data	nosql
Big data	pig
Big data	sql-h
Big data	sqoop

Topic Percentages in Documents



Topics % by Job Categories

Topics \ Job Categories	Analytics	Big Data	Business Analyst	Data Scientist
Count:	260	28	146	292
BI software	6.2%	10.7%	4.1%	3.1%
Big data	16.9%	71.4%	11.0%	50.3%
Business domain	43.5%	10.7%	45.9%	20.2%
Business intelligence	15.0%	21.4%	11.6%	19.5%
Cloud computing	2.3%	35.7%	0.7%	6.5%
Computer science	1.2%	28.6%	2.7%	24.0%
Data handling	22.7%	50.0%	28.8%	41.8%
Database	28.5%	67.9%	32.2%	53.8%
Education Grad	3.5%	7.1%	0.7%	12.3%
Education undergrad	3.1%	14.3%	4.8%	7.2%
Managerial skill	21.9%	14.3%	20.5%	17.8%
Modeling & methodology	40.8%	39.3%	29.5%	57.2%
Personal & Communication	66.9%	50.0%	67.1%	62.3%
Programming	11.5%	50.0%	4.8%	34.6%
Scripting	11.9%	42.9%	4.1%	46.9%
System analysis & design	10.4%	21.4%	24.7%	9.6%
Tools	41.5%	28.6%	27.4%	46.2%
Web Analytics	28.8%	0.0%	2.7%	4.1%

Text Rule Builder Prediction for Job Categories

Rule	Target Value ▲	Precision
google analytics	ANALYTICS	87.88%
analytics & ~python & ~data	ANALYTICS	86.67%
business acumen	ANALYTICS	87.80%
accuracy	ANALYTICS	87.91%
concise	ANALYTICS	87.88%
usage	ANALYTICS	87.16%
management & ~relational database	ANALYTICS	78.24%
tibco	ANALYTICS	78.86%
consulting	ANALYTICS	76.96%
compute & warehouse	BIG DATA	85.71%
job	BIG DATA	64.29%
process & system	BUSINESS ANALYST	86.96%
visio	BUSINESS ANALYST	90.63%
written communication	BUSINESS ANALYST	90.00%
documentation	BUSINESS ANALYST	87.76%
cost	BUSINESS ANALYST	87.27%
business analysis	BUSINESS ANALYST	85.48%
detail-orient	BUSINESS ANALYST	85.07%
government	BUSINESS ANALYST	85.71%
rapid	BUSINESS ANALYST	86.30%
enterprise	BUSINESS ANALYST	75.00%
data & ~project & model	DATA SCIENTIST	100.0%
hadoop & solver	DATA SCIENTIST	100.0%
data & phd	DATA SCIENTIST	100.0%
technique & map-reduce	DATA SCIENTIST	100.0%
science	DATA SCIENTIST	98.20%
set	DATA SCIENTIST	96.85%
cluster analysis	DATA SCIENTIST	95.14%
python	DATA SCIENTIST	89.95%
machine learn technique	DATA SCIENTIST	90.00%

Future Research

- More data collection is continuing
- Analysis of other job data (e.g., preferred requirements, preferred education, etc.)
- Trend analysis on analytic jobs requirements over time